


RESEARCH

Open Access

# *Cis* and *trans* effects differentially contribute to the evolution of promoters and enhancers



Kaia Mattioli<sup>1,2</sup>, Winona Oliveros<sup>3</sup>, Chiara Gerhardinger<sup>1</sup>, Daniel Andergassen<sup>1</sup>, Philipp G. Maass<sup>4,5</sup>, John L. Rinn<sup>6</sup> and Marta Melé<sup>3\*</sup> 

\* Correspondence: [marta.mele@bsc.es](mailto:marta.mele@bsc.es)

<sup>3</sup>Life Sciences Department, Barcelona Supercomputing Center, 08034 Barcelona, Catalonia, Spain  
Full list of author information is available at the end of the article

## Abstract

**Background:** Gene expression differences between species are driven by both *cis* and *trans* effects. Whereas *cis* effects are caused by genetic variants located on the same DNA molecule as the target gene, *trans* effects are due to genetic variants that affect diffusible elements. Previous studies have mostly assessed the impact of *cis* and *trans* effects at the gene level. However, how *cis* and *trans* effects differentially impact regulatory elements such as enhancers and promoters remains poorly understood. Here, we use massively parallel reporter assays to directly measure the transcriptional outputs of thousands of individual regulatory elements in embryonic stem cells and measure *cis* and *trans* effects between human and mouse.

**Results:** Our approach reveals that *cis* effects are widespread across transcribed regulatory elements, and the strongest *cis* effects are associated with the disruption of motifs recognized by strong transcriptional activators. Conversely, we find that *trans* effects are rare but stronger in enhancers than promoters and are associated with a subset of transcription factors that are differentially expressed between human and mouse. While we find that *cis-trans* compensation is common within promoters, we do not see evidence of widespread *cis-trans* compensation at enhancers. *Cis-trans* compensation is inversely correlated with enhancer redundancy, suggesting that such compensation may often occur across multiple enhancers.

**Conclusions:** Our results highlight differences in the mode of evolution between promoters and enhancers in complex mammalian genomes and indicate that studying the evolution of individual regulatory elements is pivotal to understand the tempo and mode of gene expression evolution.

**Keywords:** Regulatory element evolution, Gene expression evolution, Massively parallel reporter assays, *Cis* and *trans* effects



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Since it was suggested over half a century ago that changes in transcriptional regulation underlie phenotypic differences between species [1, 2], it has become clear that changes in gene expression are heritable and often play a role in the evolution of phenotypes [3, 4]. Changes in non-coding regulatory elements—including promoters and enhancers—are particularly important in driving the evolution of gene expression [5, 6]. Two primary mechanisms are responsible for the evolution of gene expression: *cis* effects and *trans* effects. *Cis* effects are due to genetic variants that are on the same DNA molecule as the target gene; for example, genetic variants located in gene promoters or enhancers that affect transcription factor (TF) binding sites. Conversely, *trans* effects are driven by diffusible elements (such as TFs) and can therefore occur anywhere in the genome. Any given gene can be subject to *cis* effects, *trans* effects, or both [7]. Characterizing the mechanisms responsible for evolutionary changes in gene expression levels remains a central goal of evolutionary biology.

Much work has assessed the contribution of *cis* and *trans* effects on the evolution of gene expression. One of the most common approaches has been to perform allele-specific RNA sequencing of two parental strains and their corresponding F1 hybrid offspring, which can separate the proportion of expression variation attributable to variants in *cis* (which show allele-specific effects in the hybrid) from expression variation attributable to variants in *trans* (which affect both hybrid alleles) [8]. These studies have assessed both intra- and inter-species variation in gene expression across a variety of taxa, including yeast [9, 10], insects [11, 12], plants [13], and mice [14]. Such hybrid methods have even been used to assess gene expression divergence between humans and mice [15], although such an approach in distantly related species is limited to examining a single artificially inserted chromosome. In general, these hybrid studies have shown a predominance of *cis* effects between species [8, 9, 11, 13, 15], with *trans* effects playing a larger role within species [10, 11, 16, 17]. Moreover, *cis* and *trans* effects were found to often occur simultaneously and affect target gene expression in opposite directions [14, 16–18]. This so-called compensation between *cis* and *trans* effects is thought to be a result of stabilizing selection on gene expression over evolutionary time [14, 16, 17]. A major limitation of these studies, however, is that while they can assign *cis* and *trans* effects to target genes, they cannot disentangle effects at individual regulatory elements. Studies on regulatory element evolution have found that the number of regulatory elements—especially enhancers—that target a gene influences the tempo and mode of gene expression evolution [5, 6]. However, to date, only small scale studies have examined how *cis* and *trans* effects drive differences in regulatory element activities across species [19, 20].

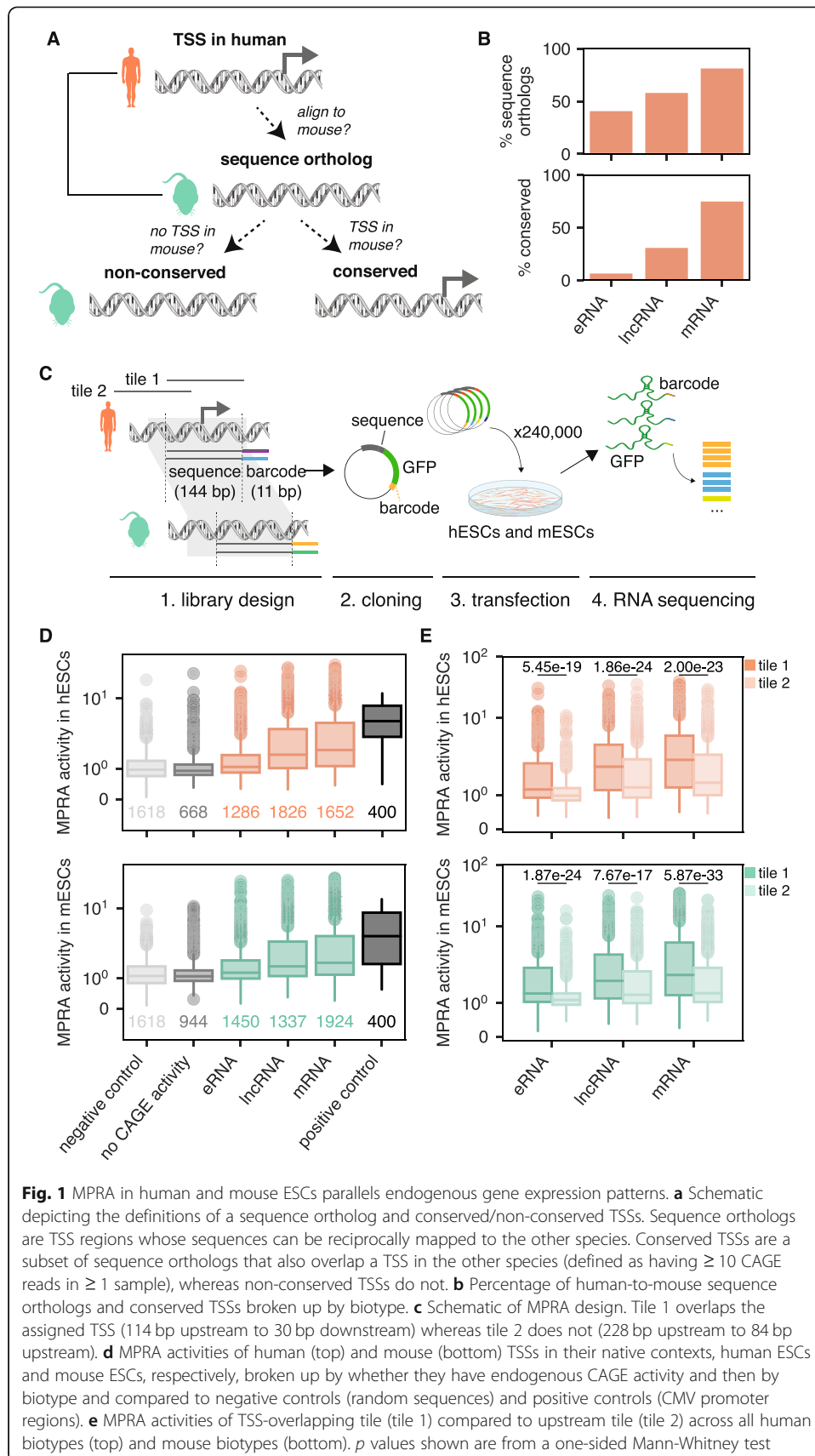
The development of massively parallel reporter assays (MPRAs) has revolutionized our ability to dissect the regulatory element code [21, 22]. Indeed, MPRAs have been used to measure regulatory element activity of thousands of sequences across tissues [23], species [20], and allelic variants [23–25]. In this work, we use MPRAs to quantitatively investigate *cis* and *trans* effects on transcriptional output across thousands of individual regulatory elements including transcribed enhancers, promoters of protein-coding genes, and promoters of long non-coding RNA (lncRNA) genes. We perform MPRAs in similar cellular environments from two mammalian species—embryonic stem cells (ESCs) from human and mouse—to perform a systematic analysis of *cis* and *trans* effects on RNA production at thousands of individual regulatory elements simultaneously.

## Results

### Designing an MPRA to measure regulatory element evolution

To investigate regulatory element evolution between human and mouse, we first defined regulatory elements in both species using a set of robust transcription start sites (TSSs) from the FANTOM5 consortium defined by Cap Analysis of Gene Expression (CAGE) sequencing [26]. We categorized these TSSs into three major biotypes: (1) eRNAs (RNAs emerging from bidirectionally transcribed enhancers that do not overlap protein-coding genes), (2) lncRNA promoters, and (3) mRNA promoters (see the “Methods” section). We then projected these TSSs onto the genome of the other species (i.e., human TSSs were projected onto the mouse genome and vice versa). We classified TSSs as “sequence orthologs” if we were able to reciprocally map the TSS between the two species. We further classified the “sequence ortholog” TSSs as conserved TSSs if the aligned region in the other species ( $\pm 50$  bp from the TSS) contained evidence of an active TSS (Fig. 1a; see the “Methods” section). As expected, the proportion of TSS that were sequence orthologs and conserved were both highest in mRNAs and lowest in eRNAs (Fig. 1b; Additional file 1: Figure S1). Despite moderate levels of sequence orthology in eRNAs and lncRNAs, both biotypes exhibited very high activity turnover, with only 7% and 31% of human eRNA TSSs and lncRNA TSSs being conserved in mouse, respectively.

To systematically assess the contribution of *cis* and *trans* effects to the evolution of thousands of regulatory elements simultaneously, we performed a massively parallel reporter assay (MPRA) and measured the transcriptional outputs of eRNA, lncRNA, and mRNA TSSs (Fig. 1c). MPRA measures the activities of designed sequences in a cell type of interest and thus enable us to test both *cis* effects (how do orthologous sequences compare within a given cellular environment) and *trans* effects (how do different cellular environments affect a given sequence). As early development is known to play a key role in evolutionary processes [27], we chose to perform the MPRA in a developmentally relevant cell type: human and mouse ESCs. Thus, we selected 3327 pairs of orthologous regulatory elements between human and mouse, all of which had endogenous activity in either human or mouse ESCs or both (Additional file 1: Figure S2; Table S1; see the “Methods” section). The full list of regulatory elements in our library can be found in Additional file 2: Table S2. To ensure that we covered all regulatory activity sequence surrounding the TSS, we designed two oligonucleotide tiles for each TSS (Fig. 1c). All told, our library included 13,533 sequences to test (Additional file 1: Table S3). To control for technical variation across sequencing measurements, each element was represented a minimum of 13 times, each time with a different barcode. We also included randomly generated sequences as negative controls (with 3 barcodes each) as well as tiled regions of the cytomegalovirus (CMV) promoter (which is known to have high activity in MPRA across diverse cell lines [23]) as positive controls (with 60 barcodes each), resulting in a final library of 181,065 unique oligonucleotides (Additional file 1: Table S4). We performed three biological replicates each in human ESCs (hESC) and mouse ESCs (mESC) and confirmed that replicates of hESC and mESC clustered separately (Additional file 1: Figures S3 and S4). We then removed barcodes with low counts, resulting in a set of 2952 regulatory sequence pairs that were well represented in our data (see the “Methods” section).



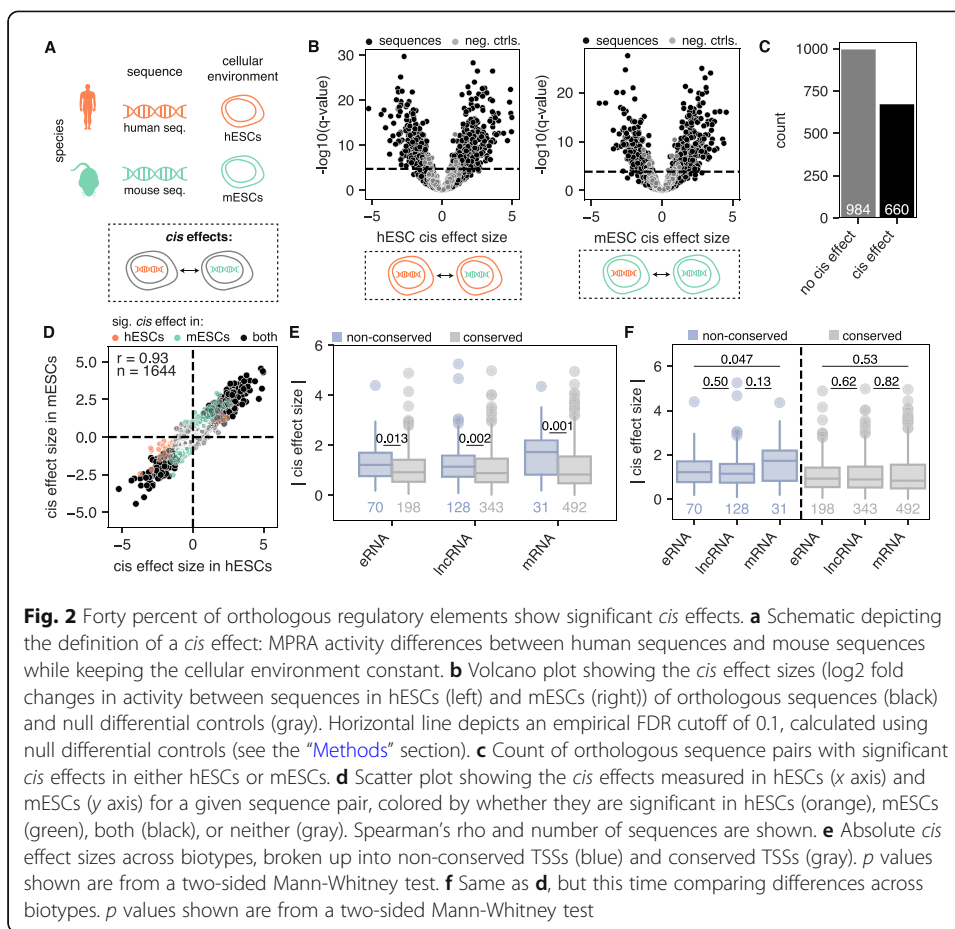
**Fig. 1** MPRA in human and mouse ESCs parallels endogenous gene expression patterns. **a** Schematic depicting the definitions of a sequence ortholog and conserved/non-conserved TSSs. Sequence orthologs are TSS regions whose sequences can be reciprocally mapped to the other species. Conserved TSSs are a subset of sequence orthologs that also overlap a TSS in the other species (defined as having  $\geq 10$  CAGE reads in  $\geq 1$  sample), whereas non-conserved TSSs do not. **b** Percentage of human-to-mouse sequence orthologs and conserved TSSs broken up by biotype. **c** Schematic of MPRA design. Tile 1 overlaps the assigned TSS (114 bp upstream to 30 bp downstream) whereas tile 2 does not (228 bp upstream to 84 bp upstream). **d** MPRA activities of human (top) and mouse (bottom) TSSs in their native contexts, human ESCs and mouse ESCs, respectively, broken up by whether they have endogenous CAGE activity and then by biotype and compared to negative controls (random sequences) and positive controls (CMV promoter regions). **e** MPRA activities of TSS-overlapping tile (tile 1) compared to upstream tile (tile 2) across all human biotypes (top) and mouse biotypes (bottom). *p* values shown are from a one-sided Mann-Whitney test

We next quantified each sequence's ability to drive transcription in the MPRA experiment—termed “MPRA activity”—using MPRAalyze [28]. Briefly, MPRAalyze uses a graphical model to estimate the rate of transcription of each sequence in the library by comparing RNA counts for each barcode to input DNA counts for each barcode. To determine whether our MPRA was able to capture true biological signal, we compared the MPRA activity of each regulatory element in its native context (human sequences in hESCs and mouse sequences in mESCs) to negative and positive control sequences (see the “Methods” section). As expected, all TSS biotypes were more active than negative controls, and eRNAs had the lowest activity across biotypes while mRNAs had the highest activity (Fig. 1d).

We then compared the activity of the annotated TSS-overlapping tiles (tile 1) to the upstream tiles (tile 2) (Fig. 1c). As expected, across all biotypes, annotated TSS-overlapping tiles were significantly more active in their native context than the upstream tiles (Fig. 1e). In 18% of regulatory element pairs, however, the upstream tile was more active than the TSS-overlapping tile in both species (Additional file 1: Figure S5), likely due to slight misannotation of the exact TSS location. Thus, while FANTOM5-defined TSSs are highly accurate, including additional upstream regions in the MPRA can help to refine core promoter locations. We therefore assigned each of the 2952 regulatory element pairs a single representative tile to use in both species: we always used the annotated TSS-overlapping tile except in those cases where the upstream tile had more activity in both species. Among those, 1644 pairs (55%) had significant MPRA activity (MPRA  $q$  value  $< 0.05$ ) in at least 1 native context. We limited all of our subsequent analyses to this set of 1644 active sequence pairs (3288 sequences total). When doing biotype-specific analyses, we focused on the set of 1262 active sequence pairs (523 mRNA, 471 lncRNA, and 268 eRNA TSSs) that could be reliably assigned to the three main biotypes (see the “Methods” section).

### ***Cis* effects are common and associated with evolutionary turnover**

Differences in regulatory element activity between species could be due to differences in DNA sequence (*cis* effects) or cellular context differences (*trans* effects) between the species or both. We decided to focus first on *cis* effects, which can be attributed to differences in DNA sequence alone. We defined *cis* effects as the MPRA activity differences between orthologous sequence pairs in the same cellular environment (Fig. 2a). To calculate *cis* effects, we used MPRAalyze to test for MPRA activity differences between pairs of orthologous regulatory elements. An advantage of using MPRAalyze is that it is able to use information from null differential controls to inform its comparative model. The ideal null differential controls are pairs of identical sequences tagged with different barcodes. We therefore leveraged our CMV tiles, each of which was attached to 60 barcodes, to create our null differential controls by down-sampling barcodes (Additional file 1: Figures S6 and S7; see the “Methods” section). As expected, orthologous regulatory element pairs had higher *cis* effect sizes than null differential controls in both hESCs and mESCs (Fig. 2b). Overall, 40% of the 1644 tested regulatory element pairs showed a significant *cis* effect in hESCs, mESCs, or both (empirical FDR  $< 0.1$ ) (Fig. 2c; see the “Methods” section). *Cis* effects were highly correlated across cell types (Fig. 2d).

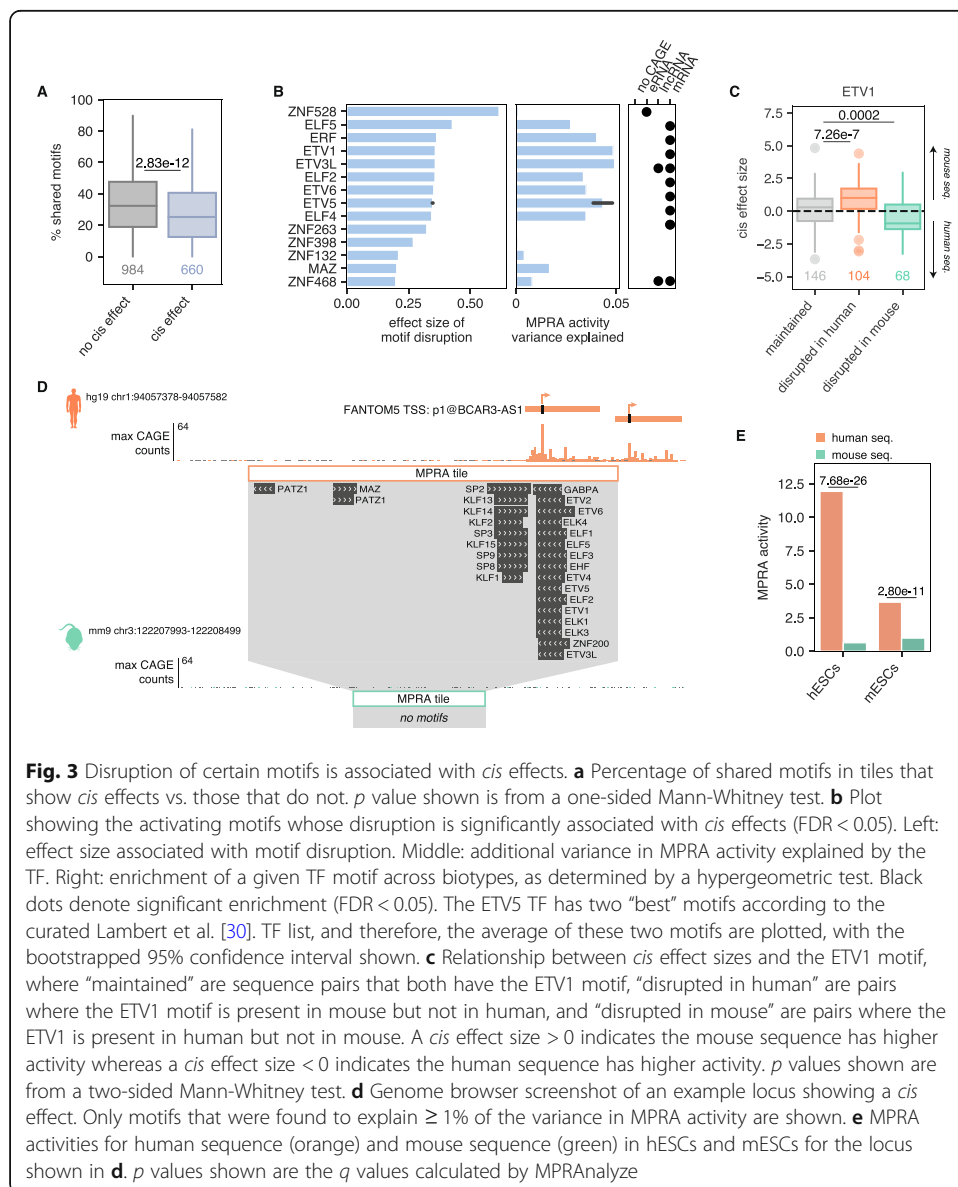


We next sought to examine how *cis* effects differ across biotypes, including conserved and non-conserved TSSs. Within each biotype, non-conserved regulatory element pairs showed significantly higher *cis* effects than conserved regulatory element pairs (Fig. 2e). We confirmed that our non-conserved pairs are *bona fide* non-conserved regulatory elements, as the non-conserved pairs we had defined (non-conserved TSSs and their orthologous sequence in the other species that lacked a TSS) had higher pairwise alignment scores than they did to the closest TSS in the other species (Additional file 1: Figure S8). Thus, these non-conserved regulatory elements are not due to misalignments between genomes. *Cis* effect sizes across biotypes were relatively uniform (Fig. 2f). However, non-conserved mRNA TSSs showed the highest *cis* effect sizes (Fig. 2f), consistent with the idea that the largest jump in activity is from mRNA TSSs—which have the highest activity out of all biotypes—to sequences without a TSS at all.

**Cis effects are associated with disruption of certain TF motifs**

*Cis* effects are often caused by disruption of motifs that are recognized by sequence-specific transcription factors (TFs). Thus, we next sought to determine the relationship between *cis* effects and TF motifs. Previous work showed that only a subset of TF motifs can be reliably associated with MPRA activity variance [29]. Thus, we selected a set of 466 motifs from TFs that are expressed in hESCs and mESCs and are associated with

MPRA activity (Additional file 1: Figure S9; Additional file 3: Table S5) either as activators or as repressors for further analysis. As expected, regulatory element pairs showing no *cis* effects shared more TF motifs than sequence pairs with significant *cis* effects (Fig. 3a), reinforcing the notion that the more TF motifs two sequences have in common, the more similar their activity levels. In addition, we found 17 individual motifs were significantly associated with *cis* effects when disrupted. The majority of these motifs were predicted activators and enriched in mRNAs (Fig. 3b); indeed, several of the strongest effect sizes could be attributed to the ETS transcription factors, including the oncogenic TF ETV1 [31] (Fig. 3c). However, we also found a subset of motifs that were predicted repressors and enriched in eRNA TSSs (Additional file 1: Figure S10). Thus, while *cis* effects can generally be attributed to the disruption of strong activating motifs, in rarer cases, *cis* effects are due to the disruption of weak repressive motifs. While this



may reflect real biological effects, it may also be due to the fact that MPRA tiles are more powered to detect activators over repressors [29, 32].

An example of a *cis* effect can be seen at the TSS for the human-specific lncRNA *BCAR3-AS1*. In human, the strongest core promoter region of *BCAR3-AS1* contains many strong activating motifs, including several ETS motifs (Fig. 3d). The orthologous region in the mouse, however, shows no CAGE activity and lacks these motifs, as there is no orthologous lncRNA in the mouse (Fig. 3d). As expected, the pairwise alignment score between the human *BCAR3-AS1* TSS region and the region shown in Fig. 3d is higher than the alignment score to the nearest mouse TSS (160.9 compared to 138.5), indicating that our MPRA tiles are correctly aligned. In our MPRA, this pair shows a significant *cis* effect: the human sequence is significantly more active than the orthologous mouse sequence in both hESCs and mESCs (Fig. 3e). Collectively, our results show that *cis* effects are common—especially in regulatory element pairs that show activity changes between species—and associated with disruption of specific TF motifs.

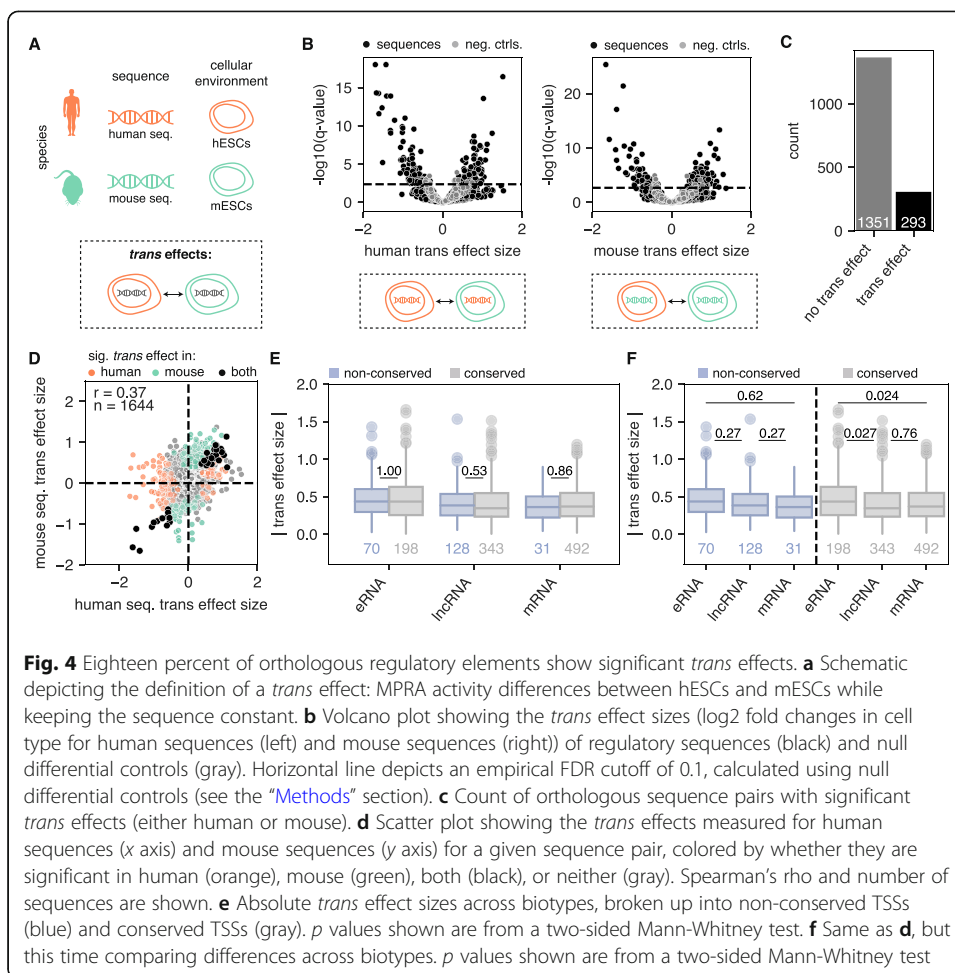
#### **Trans effects are rare and highest in eRNA TSSs**

After quantifying *cis* effects, we next sought to quantify *trans* effects. We defined *trans* effects as the difference in MPRA activity driven by differences in cellular environment alone and measured them by quantifying MPRA activity differences between hESCs and mESCs while keeping the sequence constant (Fig. 4a). As with *cis* effects, human and mouse regulatory elements showed higher *trans* effects than null differential controls (Fig. 4b). Overall, 18% of the 1644 filtered regulatory element pairs with significant activity showed a significant *trans* effect in the human sequence, the mouse sequence, or both (Fig. 4c). Compared to *cis* effect sizes, however, *trans* effect sizes were much lower. *Trans* effect sizes were also only moderately correlated across orthologous human and mouse sequences, highlighting the dominance of *cis* effects (Fig. 4d). In addition, unlike *cis* effects, we found that within each biotype, *trans* effects were similar between conserved and not conserved TSSs (Fig. 4d). While *trans* effect sizes were low in general, we found that conserved eRNA TSSs had the highest *trans* effect sizes (Fig. 4e). We speculate that this may reflect the fact that transcribed enhancers are sometimes redundant—i.e., multiple enhancers regulate the same target gene to help maintain gene expression strength [33]—and this may allow for eRNA TSSs to absorb *trans* effects at minimal fitness costs.

#### **A subset of differentially expressed TFs are associated with trans effects**

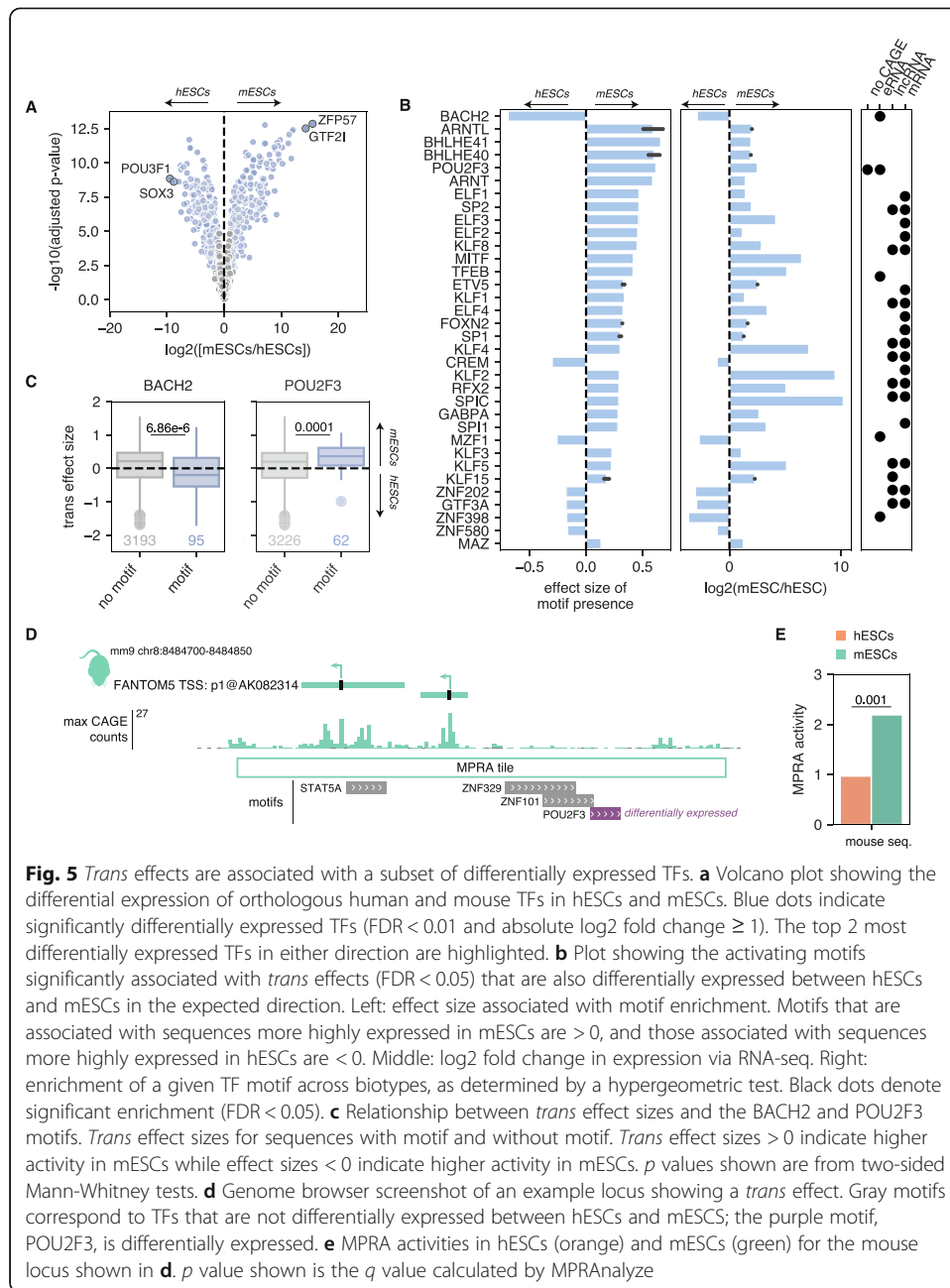
We next focused on identifying the TFs associated with the observed *trans* effects. We used a linear model to determine whether motif presence was significantly associated with *trans* effect sizes (see the “Methods”). After adjusting for multiple hypothesis testing, we found that 137 TFs (corresponding to 156 unique motifs) were significantly associated with *trans* effects. As motifs for different TFs can often be very similar to each other [30] (e.g., all POU TFs share the consensus motif ATGCAAAT), we reasoned that while we found many motifs to be significantly associated with *trans* effects, only a subset of these TFs were likely driving the *trans* effect signal. To hone in on these, we performed RNA sequencing on our hESCs and mESCs in order to find differentially expressed genes between the two species. We limited our analysis to one-to-one





orthologs between human and mouse and used a robust normalization technique (see the “Methods” section, Additional file 1: Figure S11). Of the 1032 TFs known to be one-to-one orthologs between human and mouse, 661 of these TFs were expressed in either hESCs or mESCs and 428 were significantly differentially expressed (absolute log<sub>2</sub> fold change ≥ 1 and FDR < 0.01) between hESCs and mESCs (Fig. 5a; Additional file 4: Table S6). Of the 137 TFs we found to be significantly associated with *trans* effects, 120 were one-to-one orthologs detected in our RNA-seq data. Of these, 67 were differentially expressed between hESCs and mESCs (Additional file 1: Figure S12). We reasoned that TFs likely driving *trans* effects would match in the direction of their differential expression and the direction of their *trans* effects. Of the 67 aforementioned TFs, 44 (66%) agreed in the directionality of their differential expression and *trans* effect enrichment (Fig. 5b). These included both constitutively active TFs (e.g., SP1, ARNT) as well as tissue-specific TFs (e.g., immune factor BACH2, developmental regulator POU2F3/OCT11) (Fig. 5c). We speculate that a subset of the TFs enriched in *trans* effects that are not differentially expressed may be contributing to *trans* effects through alternative mechanisms, such as evolutionary differences in TF-TF interactions.

An example of a *trans* effect associated with a differentially expressed TF can be seen at the promoter of the uncharacterized mouse lincRNA *AK082314* (Fig. 3d). This region harbors 4 motifs for 4 TFs: STAT5A, ZNF329, ZNF101, and POU2F3. Of these 4 TFs,



**Fig. 5** *Trans* effects are associated with a subset of differentially expressed TFs. **a** Volcano plot showing the differential expression of orthologous human and mouse TFs in hESCs and mESCs. Blue dots indicate significantly differentially expressed TFs (FDR < 0.01 and absolute  $\log_2$  fold change  $\geq 1$ ). The top 2 most differentially expressed TFs in either direction are highlighted. **b** Plot showing the activating motifs significantly associated with *trans* effects (FDR < 0.05) that are also differentially expressed between hESCs and mESCs in the expected direction. Left: effect size associated with motif enrichment. Motifs that are associated with sequences more highly expressed in mESCs are > 0, and those associated with sequences more highly expressed in hESCs are < 0. Middle:  $\log_2$  fold change in expression via RNA-seq. Right: enrichment of a given TF motif across biotypes, as determined by a hypergeometric test. Black dots denote significant enrichment (FDR < 0.05). **c** Relationship between *trans* effect sizes and the BACH2 and POU2F3 motifs. *Trans* effect sizes for sequences with motif and without motif. *Trans* effect sizes > 0 indicate higher activity in mESCs while effect sizes < 0 indicate higher activity in hESCs. *p* values shown are from two-sided Mann-Whitney tests. **d** Genome browser screenshot of an example locus showing a *trans* effect. Gray motifs correspond to TFs that are not differentially expressed between hESCs and mESCs; the purple motif, POU2F3, is differentially expressed. **e** MPRA activities in hESCs (orange) and mESCs (green) for the mouse locus shown in **d**. *p* value shown is the *q* value calculated by MPRAalyze

the only one that is differentially expressed between hESCs and mESCs is POU2F3, which is expressed ~ 5-fold more highly in mESCs than hESCs. Consistent with this, in our MPRA, the AK082314 promoter shows significantly higher activity in mESCs than in hESCs. Collectively, our results show that we can pinpoint a subset of TFs that may be driving *trans* effects between hESCs and mESCs.

### Co-occurrence of *cis* and *trans* effects in opposite directions is rare at eRNA TSSs

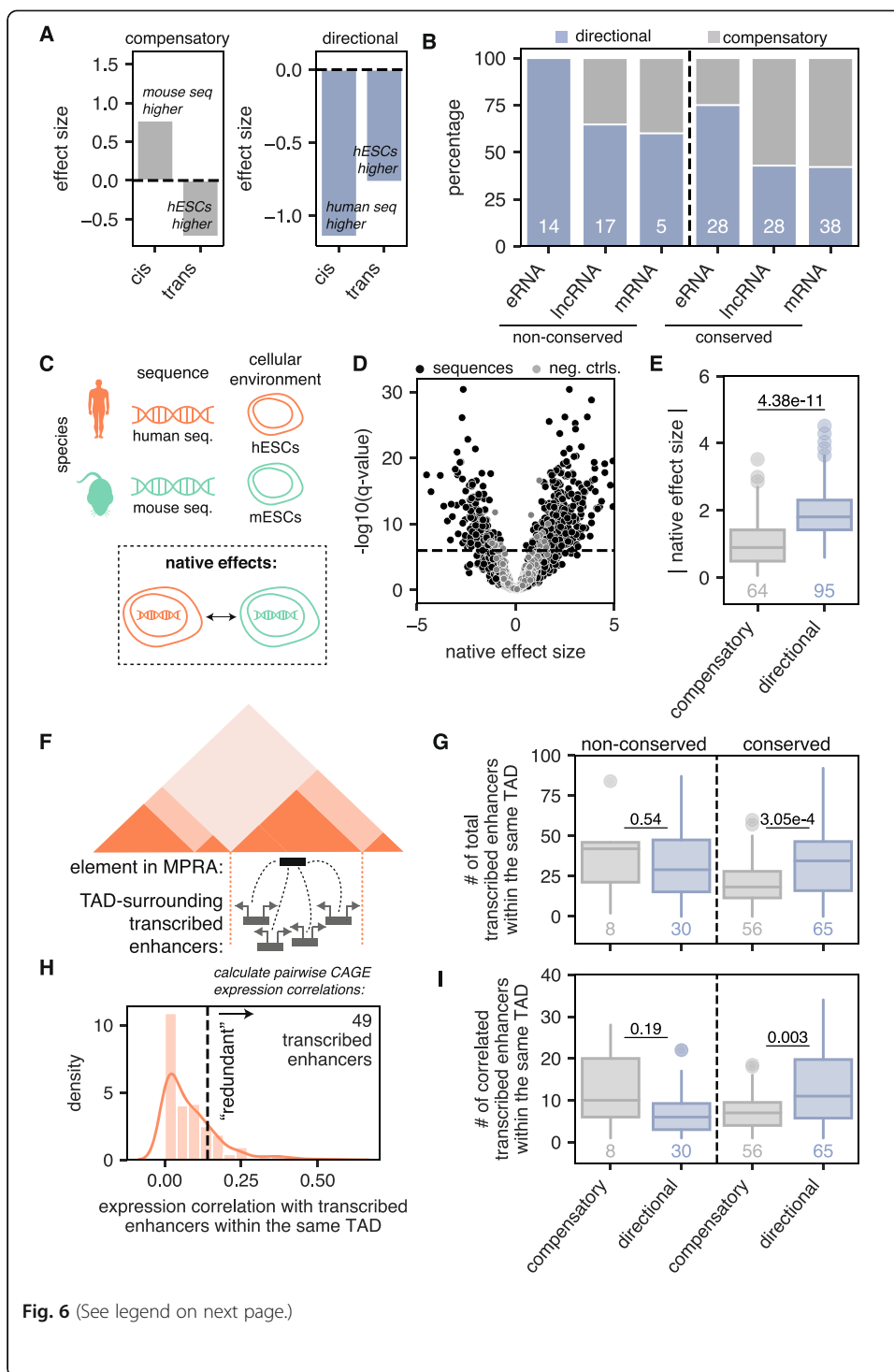
*Cis* and *trans* effects can co-occur, and previous gene-based studies have shown an excess of *cis* and *trans* effects occurring in opposite directions [14, 16–18]. These so-

called compensatory *cis-trans* effects help to stabilize gene expression over evolutionary time. It is unclear, however, whether the observed compensation between *cis* and *trans* effects occurs at the individual regulatory element level, or whether the compensation occurs primarily across different regulatory elements that regulate the same target gene [5]. We therefore sought to examine the extent of *cis-trans* compensation occurring within individual regulatory elements.

Of the 794 regulatory element pairs with either a *cis* or a *trans* effect, we found that 159 (20%) showed both *cis* and *trans* effects (odds = 2.01,  $p = 8.6 \times 10^{-8}$ , Fisher's exact test). We then determined how often the co-occurrence of *cis* and *trans* effects was compensatory (i.e., the two effects were in opposite directions—for example, the first panel of Fig. 6a depicts a regulatory element pair with a *cis* effect showing that the mouse sequence is more active, but a *trans* effect showing that the human environment results in higher activity) or “directional” (i.e., the two effects were in the same direction—for example, the second panel of Fig. 6a depicts a regulatory element pair where the *cis* and *trans* effects are both higher for the human sequence and cell type, respectively). Consistent with previous results, we found that the majority of conserved mRNA TSSs with both *cis* and *trans* effects showed compensatory *cis-trans* effects (58%, Fig. 6b). Conserved lncRNA TSSs also showed an excess of *cis-trans* compensation (57%, Fig. 6b). However, conserved eRNA TSSs showed an excess of directional *cis-trans* effects (75%, Fig. 6b), as did non-conserved TSSs (aggregated 72%, Fig. 6b). Thus, whereas mRNA and lncRNA TSS conservation is associated with *cis-trans* compensation, regulatory element turnover between human and mouse is associated with directional *cis-trans* effects. Moreover, eRNA TSSs are associated with directional effects, regardless of conservation status.

We next wondered whether the regulatory elements that show *cis-trans* compensation show evidence of stabilized activity between species. To this end, we examined the activity of orthologous regulatory element pairs in their native environments—human sequences in hESCs and mouse sequences in mESCs. We reasoned that if compensatory *cis-trans* effects stabilize regulatory element activity, we should see that regulatory elements in their native environments show virtually equal MPRA activities, so we quantified said “native effects” between orthologous regulatory element pairs (Fig. 6c, d). Indeed, regulatory element pairs showing compensatory *cis-trans* effects showed very low differences in native activity, whereas regulatory element pairs showing directional *cis-trans* effects showed large differences in native activity (Fig. 6e). Thus, quantitative regulatory element activity levels are stabilized and destabilized by compensatory and directional *cis-trans* effects, respectively.

Recent work has shown that genes regulated by larger numbers of transcribed enhancers tend to have more stable transcription throughout evolution [5]. We therefore hypothesized that perhaps regulatory elements lacking redundancy (i.e., having fewer nearby transcribed enhancers that regulate the same target gene) may show more evidence of *cis-trans* compensation than regulatory elements with higher redundancy, which show more inter-element compensation. To test this, for each regulatory element in the MPRA, we first counted the number of FANTOM5 transcribed enhancers that lied within the surrounding topologically associated domain (TAD) in either hESCs or mESCs (Fig. 6f) [34]. We found that conserved elements showing directional *cis-trans* effects were surrounded by higher numbers of transcribed enhancers (Fig. 6g).



(See figure on previous page.)

**Fig. 6** Forty percent of regulatory pairs show evidence of compensation between *cis* and *trans* effects. **a** Example of a compensatory *cis-trans* effect (left) and a directional *cis-trans* effect (right). Effect sizes > 0 indicate higher activity in the mouse sequence or cellular environment whereas effect sizes < 0 indicate higher activity in the human sequence or cellular environment. **b** Percent of regulatory element pairs across biotypes with directional *cis/trans* effects (blue) and compensatory *cis/trans* effects (gray), broken up by conservation status. Only pairs with both *cis* and *trans* effects are considered, and the total number in each group is shown. **c** Schematic showing overview of how native effects are defined. **d** Volcano plot of native effect sizes for orthologous regulatory element pairs (black) compared to null differential controls (gray). Horizontal line depicts an empirical FDR cutoff of 0.1, calculated using null differential controls (see the “Methods” section). **e** Absolute native effect sizes for sequences showing compensatory *cis-trans* effects compared to directional *cis-trans* effects. *p* value shown is from a one-sided Mann-Whitney test. **f** Schematic showing the analysis outlined in **g–i**. Darker triangles depict TADs as defined by Dixon et al. [34]. Elements in the MPRA include both gene and eRNA TSSs. **g** Number of transcribed enhancers (mean between human and mouse tiles) within the same TAD as given elements in the MPRA, broken up by conservation status. Only pairs with both *cis* and *trans* effects are considered, and the number in each group is shown. *p* values shown are from a two-sided Mann-Whitney test. **h** Example of Otsu’s method applied to threshold transcribed enhancers into those that are “redundant” (based on CAGE expression correlation) and those that are not. In this example, 49 transcribed enhancers are higher than the threshold (dashed line) and are therefore considered redundant with the element in the MPRA. **i** Number of redundant transcribed enhancers within the same TAD (mean between human and mouse tiles) as defined by Otsu’s method, broken up by conservation status. Only pairs with both *cis* and *trans* effects are considered, and the number in each group is shown. *p* values shown are from a two-sided Mann-Whitney test

Next, using a modified version of Otsu’s thresholding method [35], for each element in the MPRA, we partitioned its set of nearby transcribed enhancers into two classes: those whose CAGE activities across 1828 FANTOM5 samples are correlated with the element of interest in our MPRA (referred to as “redundant”) and those that are not (Fig. 6h, see the “Methods” section). We found that only conserved regulatory elements showing directional *cis-trans* effects were surrounded by more redundant transcribed enhancers (Fig. 6i). This observation was consistent across all conserved biotypes (Figure S13) and was robust to down-sampling after removing duplicate elements in the MPRA that happened to lie within the same TAD (Additional file 1: Figure S14).

These data show that conserved regulatory elements surrounded by higher numbers of redundant transcribed enhancers tend to show less *cis-trans* compensation than regulatory elements that are less redundant. Collectively, our results support a model whereby compensation between *cis* and *trans* effects within an individual regulatory element is more likely to occur at less redundant regulatory elements, perhaps because in these regions, there is less opportunity for inter-element compensation.

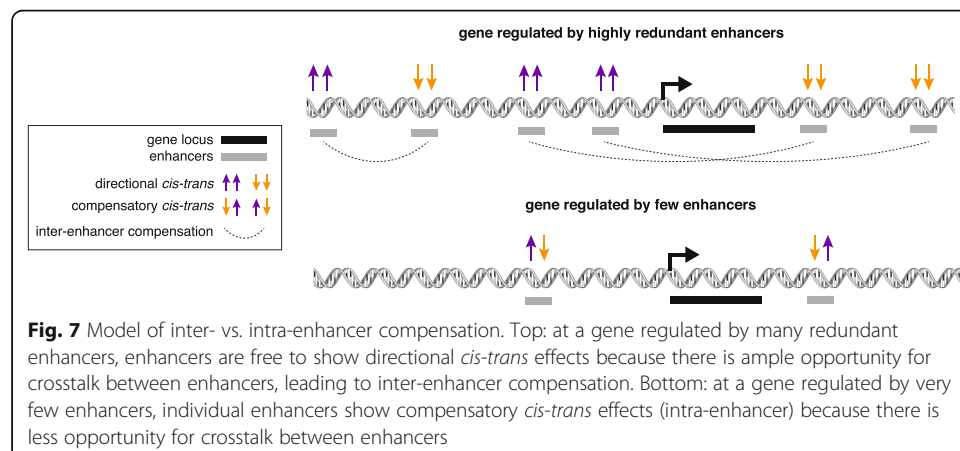
## Discussion

In this work, we sought to characterize the mode underlying the evolution of individual regulatory elements that are orthologous between human and mouse by focusing on sequences driving expression of eRNAs, lncRNAs, and mRNAs. Overall, we find that *trans* effects are less common and generally weaker than *cis* effects across all regulatory elements. These results are consistent with the prevailing model where *cis* effects preferentially accumulate between species, likely because *trans* effects result in more deleterious pleiotropic side effects that are selected against [7]. We also see differences between biotypes. While *cis* effect sizes are generally uniform across conserved eRNA, lncRNA, and mRNA TSSs (Fig. 2e), *trans* effects are highest in conserved eRNA TSSs (Fig. 4e). This suggests that the evolutionary trajectory of conserved lncRNA TSSs is more similar to that of conserved mRNA TSSs, whereas eRNA TSSs behave as a

separate group. Finally, the high resolution of our assay allowed us to identify 44 TFs that are associated with *trans* effects. Future work aimed at understanding species-level differences between human and mouse ESCs could use this set of 44 TFs as a starting point.

Previous studies have found that when *cis* and *trans* effects co-occur at the same gene, they are more often compensatory (i.e., act in different directions) than directional (i.e., act in the same direction) [14, 16–18] and are driven by stabilizing selection on transcript levels. When assessing *cis-trans* contributions at regulatory elements rather than genes, we find that conserved gene promoters—both lncRNA and mRNA TSSs—indeed show an excess of *cis-trans* compensation. Non-conserved gene promoters show less *cis-trans* compensation. Similarly, a recent publication showed that *cis-trans* compensation in TF binding was only enriched in conserved TF binding sites [36]. Interestingly, we do not find excessive *cis-trans* compensation at eRNA TSSs. In fact, both conserved and non-conserved eRNA TSSs show an enrichment of directional *cis-trans* effects (Fig. 6b). Such enrichment of directional *cis-trans* effects at eRNA TSSs may occur due to their higher redundancy compared to gene TSSs. Indeed, previous work by the FANTOM5 consortium has shown that on average, genes only have ~2 TSSs but are regulated by ~5 transcribed enhancers [33]. Moreover, recent work has shown that ensembles of redundant transcribed enhancers are often poorly conserved, despite stable expression of their target gene throughout evolution [5]. Such data is supportive of a model wherein regulatory elements can undergo evolutionary flux and compensate for one another over time. Along these lines, here, we find that regulatory elements with more redundant transcribed enhancers nearby are less likely to show compensatory *cis-trans* effects (Fig. 6i). This is consistent with the idea of inter-enhancer compensation. We propose that when regulatory elements are redundant—or have many partner regulatory elements whose activities are correlated with it—inter-enhancer compensation dominates. However, when regulatory elements are less redundant, compensation between *cis* and *trans* effects can occur at the individual regulatory element level (Fig. 7).

In this study, we sought to perform an unbiased assessment of *cis* and *trans* effects between human and mouse across a variety of biotypes. To this end, we leveraged MPRA to systematically test the contribution of *cis* and *trans* effects to the evolution of thousands of regulatory elements. An important advantage of using MPRA rather



than F1 hybrid models is that we can study the evolution of individual regulatory elements. Moreover, in MPRAs, *cis* and *trans* effects are assayed in separate experiments rather than inferred from a single hybrid (Figs. 2a and 4a). Previous work has shown that measuring *cis* and *trans* effects in the same experiment can bias the *cis* and *trans* estimates, producing spurious negative correlations between *cis* and *trans* effects and a spurious excess of *cis-trans* compensation [37]. In our study, we find no correlation between *cis* and *trans* effects (Additional file 1: Figure S15), and overall, we observe a lower rate of compensatory *cis-trans* effects compared to other studies, which is mostly driven by the lack of compensation at eRNA TSSs.

While the use of MPRAs is extremely powerful, it also has some limitations. For example, we could only study a subset of all existing regulatory elements in the human and mouse genomes. However, the sequences that we tested were carefully selected in an unbiased manner so that they would be representative of regulatory elements genome-wide. Another limitation of our approach is that we only assessed two species in one cellular background (ESCs). Although gene expression between hESCs and mESCs is similar in general, distinct differences between the two cell lines exist [38]. Moreover, whether the known differences between hESCs and mESCs are reflective of differences in isolation and culture conditions [39] or underlying species-specific biology remains controversial [40]. Future work is needed to assess whether similar patterns exist in other tissues and the extent to which these patterns may affect fundamental biological processes in a species-specific manner. Additionally, as MPRAs measure the activities of regulatory elements via transiently transfected plasmids, future work is needed to probe these elements in their native genomic context in which they evolved. Nevertheless, our work has characterized the baseline to which information from other tissues and other species can be added in order to gain a more complete picture of the evolution of regulatory elements.

## Conclusions

In summary, we find that the mode of evolution can differ at different classes of regulatory elements. Notably, we find that while compensation between *cis* and *trans* effects is common at conserved mRNA and lncRNA TSSs, it is rare at eRNA TSSs. Our results support the idea that compensation across enhancers—rather than within individual enhancers—is a widespread feature of mammalian genomes [5]. Moreover, here, we highlighted how *cis* and *trans* effects contribute differently to the transcription of eRNAs compared to the transcription of protein-coding genes and lncRNAs. While recent work shows evidence that eRNA transcription and enhancer target gene activation are linked for a subset of enhancers [41, 42], future work should focus on studying whether and how transcriptional changes at eRNAs will impact target gene expression across species. Collectively, our results underscore the importance of examining the role of individual regulatory elements in the evolution of gene expression.

## Methods

### TSS selection and biotype assignment

To assign accurate TSSs to genes, we intersected human and mouse GENCODE genes [43] (v19 in human and vM13 in mouse) with FANTOM5 TSSs [26] in both species.

Specifically, we found the closest FANTOM5 TSS (on the same strand) within  $\pm 1000$  bp of the GENCODE-annotated TSS. We classified any gene having a GENCODE `gene_type` of “protein\_coding” as an mRNA. We classified any gene included in the GENCODE `long_noncoding_RNAs` gtf file as a lncRNA, provided it showed no evidence of a conserved open reading frame (PhyloCSF [44] ORF score  $< 0$  and branch length  $< 0.1$ ). We classified any FANTOM5-annotated “robust” enhancers [33] as eRNAs and used both the sense and antisense TSS provided by FANTOM5. More details are available in Additional file 1: Supplemental Methods (TSS selection and biotype assignment section).

#### Sequence orthology assignment

To determine sequence orthologs, we first mapped human TSSs (hg19) to mouse (mm9) and vice versa using the liftOver program with the parameter `minMatch = 1`. We then reciprocally mapped the lifted-over TSSs back to their original species and required that they map to the exact same original TSS nucleotide. As FANTOM5 enhancers have two TSSs, we required that both TSSs reciprocally map in order to consider an enhancer a sequence ortholog.

#### Conserved TSS assignment

To determine conserved vs. non-conserved TSSs, we intersected the lifted-over TSSs with the maximum CAGE read coverage in that species (`ctssTotalCounts` bigwig files downloaded from the FANTOM5 data hub [45]). We determined a TSS to be conserved if the region immediately surrounding the TSS ( $\pm 50$  bp) contained  $\geq 10$  maximum CAGE reads. As enhancers have two TSSs, if either of the TSSs intersected  $\geq 10$  maximum CAGE reads, we considered it conserved.

#### MPRA sequence pair selection

We required all sequence pairs in the MPRA library to have either an annotated CAGE peak in human or mouse that is expressed above background in either hESCs or mESCs ( $\geq 0.024$  normalized counts in hESCs and  $\geq 0.022$  normalized counts in mESCs, Additional file 1: Figure S2). We included all lncRNAs (and their orthologous sequences) that met this threshold in the pool. We randomly selected the remaining biotypes in roughly equal numbers, given that they met this expression threshold. As eRNAs have two TSSs, we included both of its TSSs and both of its TSSs' orthologous sequences in the pool. Exact numbers of each biotype in the MPRA can be found in Additional file 1: Table S1, and the list of regulatory elements included in the MPRA can be found in Additional file 2: Table S2 and. More details can be found in Additional file 1: Supplemental Methods (MPRA sequence pair selection section).

#### MPRA oligonucleotide design

Each oligonucleotide we designed was 200 bp long, containing 144 bp of regulatory sequence, an 11-bp barcode, and 45 bp of sequence necessary for cloning. For each TSS selected above, we included two 144-bp tiles: one directly surrounding the TSS ( $-114/+30$  bp) and one slightly upstream of the TSS ( $-228/-84$  bp) (Additional file 1: Table S3). We then generated 1622 random 144-bp sequences to serve as negative controls.



We also tiled across the CMV promoter in 144-bp segments to create 4 positive control tiles. We assigned TSS regions 13 barcodes, random sequences 3 barcodes, and CMV sequences 60 barcodes (Additional file 1: Table S4). More details can be found in Additional file 1: Supplemental Methods (MPRA oligonucleotide design section).

#### **MPRA cloning, transfection, and sequencing**

Twist Bioscience synthesized the oligo pool, which we then cloned as previously described [23] into plasmids to generate a library of constructs where the regulatory sequence is upstream of a reporter gene (here, GFP) that is upstream of a unique barcode. We assayed the initial representation of barcodes using high-throughput DNA sequencing. We transfected these constructs into live cells and performed three biological replicates each in hESCs (HUES64 cells) and mESCs (derived from mouse blastocysts [46]) corresponding to three consecutive passages (Additional file 1: Figure S3). We isolated RNA and assayed barcode expression by high-throughput RNA sequencing. More details can be found in Additional file 1: Supplemental Methods (MPRA cloning, transfection, and sequencing section).

#### **MPRA analysis**

All code to reproduce analyses is available at [https://github.com/kmattioli/2019\\_cis\\_trans\\_MPRA](https://github.com/kmattioli/2019_cis_trans_MPRA) as well as on Zenodo at <https://doi.org/10.5281/zenodo.3862824>.

#### **Quantifying MPRA activity**

After trimming and quality filtering DNA and RNA reads, we mapped exact matches to known barcodes and 10 upstream constant nucleotides of GFP. We only measured sequences that had at least 50% of their barcodes represented at  $\geq 10$  counts in the input DNA library. We used the R package MPRAalyze [28] to quantify MPRA activities for each sequence in each condition using the program's "quantification" mode. We used our randomly generated sequences as the background null distribution, as the majority of these sequences should not induce transcription. More details can be found in Additional file 1: Supplemental Methods (quantifying MPRA activities section).

#### **Calculating differential MPRA activity**

After quantifying MPRA activity and assigning 1 tile to each sequence pair (Additional file 1: Figure S5), we used MPRAalyze [28] to perform differential activity analyses using the program's "comparison" mode. In comparison mode, as the null hypothesis is not the lack of transcription but the lack of differential transcription, we used down-sampled barcodes corresponding to identical CMV sequences as the background null distribution. In each of the 5 models (*cis* effects in hESCs, *cis* effects in mESCs, *trans* effects of mouse sequences, *trans* effects of human sequences, and native effects), we tested whether the full model was a better fit than an intercept-only model using a likelihood ratio test. More details can be found in Additional file 1: Supplemental Methods (calculating differential MPRA activity section).

### Calling significant differential effects

We considered sequences to have significant *cis*, *trans*, or native effects if the  $q$  value calculated by MPRAalyze was less than the  $q$  value that resulted in < 10% of negative controls being called significant, which is effectively an empirical FDR of 0.1 (Additional file 1: Figure S6). We also required effect sizes to be higher than the minimum significant null differential control effect size (Additional file 1: Figure S7). We assigned each sequence pair one *cis* and *trans* effect size: we used the maximum *cis* or *trans* effect size between the two models (hESCs/mESCs for *cis* and human/mouse for *trans*) unless the effect was only significant in one model, in which case used the corresponding significant effect size. More details can be found in Additional file 1: Supplemental Methods (Calling significant differential effects section).

### Motif mapping

We used a curated list of human TFs defined by Lambert et al. [30]. We then used the CisBP [47] position-weight matrices designated by Lambert et al. to be the “best” motifs for each of these TFs. In total, this list contained 1360 motifs corresponding to 1104 unique TFs. We mapped these motifs in both human sequences and mouse sequences using the FIMO program from the MEME suite with default parameters [48].

### Finding motifs predictive of MPRA activity

For each motif, we fit a linear model to mean MPRA activity across all sequences as follows:

$$\text{mean}(\text{MPRA activity}) \sim \text{GC content} + \text{CpG content} + \text{motif present}$$

and determined whether the binary *motif present* indicator explained significantly more of the variance than a reduced model without the indicator using a likelihood ratio test (Additional file 3: Table S5). We used the Python statsmodels [49] package to run all linear models. More details can be found in Additional file 1: Supplemental Methods (Finding motifs predictive of MPRA activity section).

### Finding motifs associated with *cis* and *trans* effects

For each motif, we fit a linear model to absolute *cis* effect sizes across all sequence pairs as follows:

$$|\text{cis effect size}| \sim \text{mean}(\text{GC}) + \text{mean}(\text{CpG}) + |\Delta(\text{GC})| + |\Delta(\text{CpG})| + \text{motif disrupted}$$

and determined whether the *motif disrupted* parameter (indicating whether a motif was present in only one of the two paired sequences) was significant (FDR < 0.05).

For each motif, we fit a linear model to *trans* effect sizes across all sequences as follows:

$$\text{trans effect size} \sim \text{GC} + \text{CpG} + \text{motif present}$$

and determined whether the *motif present* parameter was significant (FDR < 0.05). More details available in Additional file 1: Supplemental Methods (Finding motifs associated with *cis* and *trans* effects sections).

### RNA-seq of hESCs and mESCs

We sequenced both untransfected and transfected hESCs and mESCs. We extracted RNA from TRIzol using standard protocols and used the Illumina TruSeq kit (non-stranded) to create polyA+ libraries from total RNA. We measured library concentration using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific), and ran all of the libraries on a Bioanalyzer (Agilent) to assess purity and fragment size, and sequenced on a HiSeq 2500 at Harvard University's Bauer Sequencing Core (75 bp paired end).

### RNA-seq analysis

We aligned reads to either hg19 or mm10 using Hisat2 [50]. We used FeatureCounts to count reads aligning to genes in either GENCODE v25 (human) or GENCODE vM13 (mouse) [51]. We downloaded orthologous genes between human and mouse from Ensembl (version 96) [52] and removed any orthologs classified as “many-to-many.” We normalized gene expression values using the trimmed mean of  $M$  values (TMM) normalization method in edgeR [53], similar to previous cross-species comparisons [54, 55]. Briefly, TMM normalization re-scales samples relative to each other under the assumption that most genes are not differentially expressed [53]. This assumption is valid when comparing human and mouse ESC expression, because even between more distant mammals—such as humans and opossums—gene expression is tightly correlated [56]. To find differentially expressed genes, we used the edgeR-limma pipeline [53] (filtering out any genes with normalized cpm < 1) to model paired samples (transfected and untransfected) and control for transfection status. For plotting purposes, we quantified gene expression in tpm units in each transfected sample using DESeq2 [57].

### Defining redundant enhancers

To find redundant enhancers, we first downloaded the CAGE-seq expression values for all enhancer and promoter TSSs from the FANTOM5 portal [45]. For every element in the MPRA, we then found all enhancers within the same TAD defined in either hESCs or mESCs by Dixon et al. [34]. We then calculated the Pearson correlation coefficient of CAGE-seq expression (log-transformed) between each element in the MPRA and its TAD-surrounding enhancers. We then used a modified version of Otsu's method [35] to threshold enhancers at the appropriate correlation cutoff. Otsu's method is typically used to automatically threshold bimodal grayscale images into “black” pixels and “white” pixels by creating a histogram of pixel values which range from 1 to 255. We created an analogous histogram of correlation coefficients using 100 bins between 0 and 1 (the range of correlation coefficients). We considered the number of “redundant” enhancers to be the number of TAD-surrounding enhancers above the Otsu cutoff.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02110-3>.

**Additional file 1.** Supplemental Methods, Supplemental Figures S1-S15, and Supplemental Tables S1, S3, and S4.

**Additional file 2: Supplemental Table S2.** list of regulatory elements included in the MPRA.

**Additional file 3: Supplemental Table S5.** TF motifs and their effects on MPRA activity.

**Additional file 4: Supplemental Table S6.** orthologous TF expression in hESCs and mESCs.

**Additional file 5.** Review history.

### Acknowledgements

We thank Veronika Akopian for help with HUES64 cell culture, Abigail Groff for providing mESCs, Jordan Lewandowski for help with mESC cell culture, and Martha Bulyk for helpful discussions.

### Review history

The review history is available as Additional file 5.

### Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Authors' contributions

K.M. and M.M. designed the project and wrote the manuscript. K.M. designed the oligonucleotide libraries and performed all MPRA computational analyses. W.O. analyzed the FANTOM5 CAGE data and performed the RNA-seq analysis. C.G. and P.G.M. performed the MPRA experiments. C.G. performed the RNA-seq. D.A. assisted with the cell culture. J.L.R. contributed to the project design and discussion. All authors have read and approved the manuscript for publication.

### Authors' information

@kaia\_mattioli (Kaia Mattioli); @OliverosWinona (Winona Oliveros); @DAndergassen (Daniel Andergassen); @NucleomeHunter (Philipp G. Maass); @Noncodarnia (John L. Rinn); @marta\_mele\_m (Marta Melé).

### Funding

K.M. was a National Science Foundation Graduate Research Fellow under grant no. DGE1144152 during the majority of the project. M.M. was a Gilead Fellow of the Life Sciences Research Foundation during part of the project and is currently supported by the Spanish Ministry of Science and Innovation with a Ramon y Cajal grant (RYC-2017-22249). J.L.R. is an HHMI faculty scholar.

### Availability of data and materials

The MPRA sequencing data and genomic RNA-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE140574 [58]. All scripts required to reproduce this work are available on GitHub at [https://github.com/kmattioli/2019\\_\\_cis\\_trans\\_MPRA](https://github.com/kmattioli/2019__cis_trans_MPRA) and are archived on Zenodo at <https://doi.org/10.5281/zenodo.3862824> [59]. The code is open source and released under the MIT License.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

<sup>1</sup>Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA. <sup>2</sup>Department of Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA. <sup>3</sup>Life Sciences Department, Barcelona Supercomputing Center, 08034 Barcelona, Catalonia, Spain. <sup>4</sup>Genetics and Genome Biology Program, SickKids Research Institute, Toronto, ON M5G 0A4, Canada. <sup>5</sup>Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A1, Canada. <sup>6</sup>Department of Biochemistry, University of Colorado, BioFrontiers Institute, Boulder, CO 80301, USA.

Received: 5 December 2019 Accepted: 16 July 2020

Published online: 20 August 2020

### References

1. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. *Science*. 1969;165(3891):349–57.
2. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science*. 1975;188(4184):107–16.
3. Stern DL, Orgogozo V. The loci of evolution: how predictable is genetic evolution? *Evolution*. 2008;62(9):2155–77.
4. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet*. 2012;13(7):505–16.
5. Danko CG, Choate LA, Marks BA, Rice EJ, Wang Z, Chu T, et al. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nat Ecol Evol*. 2018;2(3):537–48.
6. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol*. 2018;2(1):152–63.
7. Signor SA, Nuzhdin SV. The evolution of gene expression in cis and trans. *Trends Genet*. 2018;34(7):532–44.
8. Wittkopp PJ, Haerum BK, Clark AG. Evolutionary changes in cis and trans gene regulation. *Nature*. 2004;430(6995):85–8.
9. Tirosh I, Reikhav S, Levy AA, Barkai N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*. 2009;324(5927):659–62.
10. Emerson JJ, Hsieh L-C, Sung H-M, Wang T-Y, Huang C-J, Lu H-H, et al. Natural selection on cis and trans regulation in yeasts. *Genome Res*. 2010;20(6):826–36.

11. Wittkopp PJ, Haerum BK, Clark AG. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat Genet.* 2008;40(3):346–50.
12. Osada N, Miyagi R, Takahashi A. Cis- and trans-regulatory effects on gene expression in a natural population of *Drosophila melanogaster*. *Genetics.* 2017;206(4):2139–48.
13. Shi X, Ng DW-K, Zhang C, Comai L, Ye W, Chen ZJ. Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat Commun.* 2012;3:950.
14. Goncalves A, Leigh-Brown S, Thybert D, Stefflova K, Turro E, Flicek P, et al. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* 2012;22(12):2376–84.
15. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, et al. Species-specific transcription in mice carrying human chromosome 21. *Science.* 2008;322(5900):434–8.
16. Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. Tempo and mode of regulatory evolution in *Drosophila*. *Genome Res.* 2014;24(5):797–808.
17. Metzger BPH, Wittkopp PJ, Coolon JD. Evolutionary dynamics of regulatory changes underlying gene expression divergence among *Saccharomyces* species. *Genome Biol Evol.* 2017;9(4):843–54.
18. Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics.* 2005;171(4):1813–22.
19. Gordon KL, Ruvinsky I. Tempo and mode in evolution of transcriptional regulation. *PLoS Genet.* 2012;8(1):e1002432.
20. Ryu H, Inoue F, Whalen S, Williams A, Kircher M, Martin B, et al. Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *BioRxiv.* 2018. <https://doi.org/10.1101/256313>.
21. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol.* 2012;30(3):265–70.
22. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol.* 2012;30(3):271–7.
23. Mattioli K, Volders P-J, Gerhardinger C, Lee JC, Maass PG, Melé M, et al. High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res.* 2019;29(3):344–55.
24. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell.* 2016;165(6):1530–45.
25. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell.* 2016;165(6):1519–29.
26. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. *Nature.* 2014;507(7493):462–70.
27. Hall BK. Evolutionary developmental biology (Evo-Devo): past, present, and future. *Evo Edu Outreach.* 2012;5(2):184–93.
28. Ashuaich T, Fischer DS, Kreimer A, Ahituv N, Theis FJ, Yosef N. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* 2019;20(1):183.
29. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, et al. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol.* 2016;34(11):1180–90.
30. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. *Cell.* 2018;172(4):650–65.
31. Jané-Valbuena J, Widlund HR, Perner S, Johnson LA, Dibner AC, Lin WM, et al. An oncogenic role for ETV1 in melanoma. *Cancer Res.* 2010;70(5):2075–84.
32. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One.* 2019;14(6):e0218073.
33. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. *Nature.* 2014;507(7493):455–61.
34. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012;485(7398):376–80.
35. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern.* 1979;9(1):62–6.
36. Wong ES, Schmitt BM, Kazachenka A, Thybert D, Redmond A, Connor F, et al. Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nat Commun.* 2017;8(1):1092.
37. Fraser HB. Improving estimates of compensatory cis-trans regulatory divergence. *Trends Genet.* 2019;35(1):3–5.
38. Ginis I, Luo Y, Miura T, Thies S, Brandenberger R, Gerecht-Nir S, et al. Differences between human and mouse embryonic stem cells. *Dev Biol.* 2004;269(2):360–80.
39. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature.* 2007;448(7150):196–9.
40. Takahashi S, Kobayashi S, Hiratani I. Epigenetic differences between naïve and primed pluripotent stem cells. *Cell Mol Life Sci.* 2018;75(7):1191–203.
41. Gu B, Swigut T, Spencley A, Bauer MR, Chung M, Meyer T, et al. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science.* 2018;359(6379):1050–5.
42. Fitz J, Neumann T, Steininger M, Wiedemann E-M, Garcia AC, Athanasiadis A, et al. Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat Genet.* 2020;52(5):505–15.
43. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D766–73.
44. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics.* 2011;27(13):i275–82.
45. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* 2015;16:22.
46. Groff AF, Barutcu AR, Lewandowski JP, Rinn JL. Enhancers in the Peril lincRNA locus regulate distant but not local genes. *Genome Biol.* 2018;19(1):219.
47. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell.* 2014;158(6):1431–43.

48. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017–8.
49. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python [Internet]; 2010. p. 57. [cited 2019 Nov 14]. Available from: <http://statsmodels.sourceforge.net/>.
50. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–60.
51. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–30.
52. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. *Nucleic Acids Res*. 2017;45(D1):D635–42.
53. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
54. Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol*. 2015;16:287.
55. Ramsey SA. A method for cross-species visualization and analysis of RNA-sequence data. *Methods Mol Biol*. 2018;1702:291–305.
56. Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, et al. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res*. 2019;29(1):53–63.
57. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
58. Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, et al. Cis and trans effects differentially contribute to the evolution of promoters and enhancers. GSE140574. *Gene Expression Omnibus*. 2020. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140574>.
59. Mattioli K, Oliveros W, Gerhardinger C, Maass PG, Rinn JL, Melé M. Cis and trans effects differentially contribute to the evolution of promoters and enhancers. GitHub. 2020. <https://doi.org/10.5281/zenodo.3862824>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

